MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0145727 1

ORNL-6328

ornl

OAK RIDGE NATIONAL LABORATORY

MARTIN MARIETTA

Bayesian Variable Selection in Regression

T. J. Mitchell
J. J. Beauchamp

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
CIRCULATION SECTION
4500N ROOM 175
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this report, send in name with report and the library will arrange a loan.

OPERATED BY
MARTIN MARIETTA ENERGY SYSTEMS, INC.
FOR THE UNITED STATES
DEPARTMENT OF ENERGY

Printed in the United States of America. Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Virginia 22161
NTIS price codes—Printed Copy: A03; Microfiche A01

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ORNL-6328

Engineering Physics and Mathematics Division

Mathematical Sciences Section

BAYESIAN VARIABLE SELECTION IN REGRESSION

T. J. Mitchell
J. J. Beauchamp

Date Published: January 1987

This work was supported by the
Applied Mathematical Sciences Research Program
U.S. Department of Energy
Office of Energy Research

MARTIN MARIETTA ENERGY SYSTEMS LIBRARIES
3 4456 0145727 1

Prepared by the
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831
operated by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
for the
U.S. DEPARTMENT OF ENERGY
under Contract No. DE-AC05-84OR21400

TABLE OF CONTENTS

ABSTRACT 1

1. INTRODUCTION 1

2. METHOD 2

2.1 The Data 2
2.2 The Prior Distribution 3
2.3 The Posterior Distribution 4
2.4 Evaluation of γ 6

3. EXAMPLES 8

3.1 Background 8
3.2 Example 1: The Real Data 8
3.3 Example 2: "Important" and "Unimportant" Predictors 9
3.4 Example 3: Collinearity 11

4. DISCUSSION 13

4.1 Centering and Scaling 13
4.2 Interpretation of Bayesian Probabilities 16
4.3 A Couple of Paradoxes 17
4.4 Other Methods 18

APPENDIX 21

ACKNOWLEDGEMENT 24

REFERENCES 25

BAYESIAN VARIABLE SELECTION IN REGRESSION

T. J. Mitchell
J. J. Beauchamp

ABSTRACT

This paper is concerned with the selection of subsets of "predictor" variables in a linear regression model for the prediction of a "dependent" variable. We take a Bayesian approach and assign a probability distribution to the dependent variable through a specification of prior distributions for the unknown parameters in the regression model. The appropriate posterior probabilities are derived for each submodel and methods are proposed for evaluating the family of prior distributions. Examples are given that show the application of the Bayesian methodology.

1. INTRODUCTION

This paper is concerned with the prediction of an unknown "dependent" variable given known values of k "predictor" variables based on a statistical analysis of n cases in which all variables are measured. This is often done using linear regression methods, which are based on the statistical model:

$$y = \sum_{j=1}^{k} \beta_j x_j + \epsilon, \tag{1.1}$$

where y corresponds to the dependent variable, x_j corresponds to the jth predictor, β_j is the jth regression coefficient, and ϵ is a "random error". The regression coefficients are model parameters whose values are constant over all cases, while the value of ϵ varies randomly from case to case. It will usually be desirable to include a constant term in the model; in this case define $x_1 = 1$. We shall assume here that the remaining predictors represent distinct observables and that none of them are defined as functions of others. The reason for this limitation is that our primary method will be based on the specification of independent and identical prior distributions for the regression coefficients that are subject to deletion from the model. We think this is reasonable when there are no functional dependencies among terms and when each predictor is suitably scaled. (See Section 4.1.) However, it may not be reasonable when there are functional dependencies. We have not investigated this case sufficiently to determine what change in our method would be required.

At some point during the analysis, one may be interested in the possibility of omitting some predictors from the model. The search for a "best" submodel (or set of submodels) is called *variable selection* or *subset selection*. Some reasons for undertaking this search are: (1) to express the relationship between y and the predictors as simply as possible; (2) to reduce future cost of prediction; (3) to identify "important" and "negligible" predictors; or (4) to increase the precision of statistical estimates and predictions.

There are numerous classical approaches to variable selection. (For some of the most popular, see Hocking (1976) or Chapter 6 of Draper and Smith (1981).) These generally are based on sequences of hypothesis tests (e.g., stepwise regression) or on estimates of some type of mean squared error (e.g., Allen (1971a, 1971b), Mallows (1973)).

Here we take a Bayesian approach. We shall assign a probability distribution (the "prior distribution") to the β's and ϵ's, and hence to the y's, through (1.1). Following Box and Meyer (1986), we require that the distribution of each β_j that is a candidate for exclusion must include a discrete probability mass at the point $\beta_j = 0$. Posterior probabilities are computed as usual using Bayes' Theorem. We shall be concerned primarily with the posterior probabilities of the various submodels, where each submodel is defined by the event that a specific subset of the β's has the value 0 and the remaining β's do not. These can then be used to generate other probabilities of interest in variable selection, e.g., $P(\beta_j = 0)$ for each j. Other Bayesian approaches have been developed in the context of model discrimination; we shall discuss these in Section 4.4.

Frequentist and Bayesian approaches alike usually start with the assumption that the observations on y, given the value of the predictor vector x, are generated by a mechanism that produces stochastic data whose frequency distribution is based on the distribution of ϵ in (1.1). The elements of the parameter vector β are then regarded as "real" but unknown quantities. If one takes this view, then the placement of a discrete prior probability mass at $\beta_j = 0$ is a recognition of the possibility that β_j might be precisely zero, or at least close enough to zero to make this type of prior a reasonable approximation to one's "true" prior. In most applications, however, it is more appropriate to regard the model solely as a predictive device. As such, its parameters are artificial components of that device, not properties of the real world that are to be estimated. Then β_j is whatever one chooses it to be, and the type of prior we consider allows for the possibility that it is chosen to be zero.

In Section 2, we describe our family of Bayesian models, present the appropriate posterior distributions, and propose methods for evaluating the prior distributions within the family. Examples of the application of these methods are given in Section 3. Section 4 considers the relationship of this paper to previous work on this subject, and also discusses philosophical and technical issues that we found convenient to defer until the end. The Appendix provides some useful formulas and computational details.

2. METHOD

2.1 THE DATA

The observed data on the predictors are contained in the $n \times k$ matrix X, where the ith row of X contains the values of the predictors in the ith case. We shall assume throughout that the rank of X is k, and that $k < n$. All probability statements in this paper are implicitly conditioned on X. The observed data on the dependent variable y are contained in the n-vector $\boldsymbol{y}$.

There are 2^k possible submodels of the model (1.1), where each submodel excludes a particular subset of the predictors and includes the rest. We shall denote by X_m the $n \times k_m$ matrix consisting of those columns of X that correspond to the predictors included in the mth submodel A_m for $m = 1,2,\ldots,2^k$. The least-squares estimate of the corresponding parameter vector $\boldsymbol{\beta}_m$ is given by:

$$\hat{\beta}_m = (X_m{}'X_m)^{-1}X_m{}'y, \tag{2.1}$$

and the residual sum of squares for this fit is given by:

$$S_m^2 = (y - X_m\hat{\beta}_m)'(y - X_m\hat{\beta}_m) \tag{2.2}$$

2.2 THE PRIOR DISTRIBUTION

For each case, we shall assign to ϵ in (1.1) a normal distribution with mean 0 and variance σ^2, and we shall take the ϵ's in distinct cases to be independent. Given β and σ^2, then, the n-vector y has a multivariate normal distribution with mean $X\beta$ and covariance matrix $\sigma^2 I$. In Bayesian regression, β and (usually) σ are also random variables, and different Bayesian techniques are characterized by the choice of prior distributions for these parameters. Here we assign σ the standard "noninformative" prior, under which $ln(\sigma)$ is locally uniform. That is, $ln(\sigma)$ will be uniformly distributed between $-ln(\sigma_0)$ and $ln(\sigma_0)$, where σ_0 is very large. We further assume that the prior distribution of β is independent of σ and that the individual β's are mutually independent, each having a "spike and slab" distribution. That is, β_j is uniformly distributed between the two limits $-f_j$ and f_j, except for a bit of probability mass concentrated at 0 if x_j is vulnerable to deletion. Formally,

$$P(\beta_j = 0) = h_{0j}, \tag{2.3a}$$

$$P(\beta_j < b\ ,\ \beta_j \neq 0) = (b + f_j)h_{1j},\ -f_j < b < f_j, \tag{2.3b}$$

$$P(|\beta_j| > f_j) = 0, \tag{2.3c}$$

where $h_{0j} > 0$, $h_{1j} > 0$, and $h_{0j} + 2h_{1j}f_j = 1$. We shall take f_j and γ_j as the parameters of this distribution, where

$$\gamma_j = h_{0j}/h_{1j} = 2h_{0j}f_j/(1-h_{0j}) \tag{2.4}$$

i.e., γ_j is the height of the spike divided by the height of the slab. To exempt certain terms (e.g., the constant term) from deletion from the model, set the corresponding h_{0j} (and hence γ_j) equal to 0.

With the above model specification, our prior distribution over the submodels is :

$$P(A_m) = \prod_{\bar{J}} h_{0j} \prod_{J} (2f_j h_{1j}) = \prod_{\bar{J}} \gamma_j \prod_{J} (2f_j) \prod_{j} (\gamma_j + 2f_j)^{-1} \tag{2.5}$$

where J is the set of subscripts corresponding to the terms that are included in A_m, and $\bar{J}$ is the set of subscripts for the terms omitted from A_m.

We are interested here in taking f_j to be very large for all j, to specify prior impartiality about the value of β_j in those models in which x_j appears.

2.3 THE POSTERIOR DISTRIBUTION

Under the Bayesian model specified above, the probability density function of y given A_m, β_m, and σ is:

$$p(y \mid A_m, \beta_m, \sigma) = (2\pi)^{-n/2}\sigma^{-n} exp[-\frac{1}{2\sigma^2}(S_m{}^2 + (\beta_m - \hat{\beta}_m)'X_m'X_m(\beta_m - \hat{\beta}_m))]. \quad (2.6)$$

(We shall use a lower case p generically to mean "the density function of.")

If we multiply (2.6) by $p(\beta_m \mid A_m, \sigma)$, which is $\prod_J (2f_j)^{-1}$ over the region of positive probability, and integrate over β_m, we obtain

$$p(y \mid A_m, \sigma) = \prod_J (2f_j)^{-1}(2\pi)^{-(n-k_m)/2} \mid X_m'X_m \mid^{-1/2} \sigma^{-(n-k_m)} exp[-S_m{}^2/(2\sigma^2)] \quad (2.7)$$

where k_m is the number of terms in submodel A_m. To obtain (2.7), we assume that each f_j is large enough so that, for all $\sigma < \sigma_0$, all the integrals from $-f_j$ to f_j can be replaced by integrals from minus infinity to infinity with arbitrarily small error. Now multiply (2.7) by $p(\sigma \mid A_m)$, which is proportional to σ^{-1}, and integrate over σ to obtain

$$p(y \mid A_m) = \quad (2.8)$$
$$(2ln(\sigma_0))^{-1} \prod_J (2f_j)^{-1} \pi^{-(n-k_m)/2} (1/2)\Gamma(\frac{n-k_m}{2}) \mid X_m'X_m \mid^{-1/2} (S_m{}^2)^{-(\frac{n-k_m}{2})}.$$

In deriving (2.8), we assume that σ_0 is large enough so that the integral from σ_0^{-1} to σ_0 can be replaced by the integral from 0 to infinity with arbitrarily small error. Now (2.8) may be multiplied by $P(A_m)$ from (2.5) to obtain $p(y, A_m)$, from which we obtain:

$$P(A_m \mid y) = g \prod_{\bar{J}} \gamma_j \pi^{k_m/2} \Gamma(\frac{n-k_m}{2}) \mid X_m'X_m \mid^{-1/2} (S_m{}^2)^{-(\frac{n-k_m}{2})}, \quad (2.9)$$

where g is a normalizing constant. (For the "empty" or "null" model, replace the determinant by 1.)

To use (2.9), one needs to specify all the γ_j's. This can be done using (2.4) if all the parameters of the prior distribution (2.3) are prescribed. Here we propose a less direct approach, which requires no prior information from the user regarding the relative effects of the predictors. A modification of this approach, which allows the user more control over the prior distribution, is suggested briefly at the end of Section 4.2.

We suppose that the β's corresponding to the vulnerable predictors have identical prior distributions. (Whether or not this is a sensible choice is bound to the question of how to scale the predictors. We shall discuss this in Section 4.1. For now, assume that all vulnerable predictors have been "suitably scaled.") For all submodels having positive prior probability, (2.9) becomes

$$P(A_m \mid y) = g' \gamma^{-k_m} \pi^{k_m/2} \Gamma(\frac{n-k_m}{2}) \mid X_m'X_m \mid^{-1/2} (S_m{}^2)^{-(\frac{n-k_m}{2})}, \quad (2.10)$$

where g' is another normalizing constant, and γ is the common value of all the positive γ_j's. We consider γ to be a measure of one's prior inclination to omit predictors from the model. We treat it here as an adjustable parameter of the Bayes model, i.e., we shall not assign a distribution to it, nor do we intend it to be chosen *a priori*. Methods for assessing values of γ will be presented in Section 2.4 and illustrated in the examples of Section 3.

Equation (2.10) can also be written:

$$P(A_m \mid y) = w_m / \sum_{u=1}^{2^k} w_u \tag{2.11}$$

where the logarithm of any positive "weight" w_m is given by

$$ln(w_m) = \tag{2.12}$$
$$k_m(-ln(\gamma)+\tfrac{1}{2}ln(\pi))+ln(\Gamma(\tfrac{1}{2}(n-k_m)))-\tfrac{1}{2}ln\,|X_m'X_m|-\tfrac{1}{2}(n-k_m)ln(S_m^2).$$

Remark 1. Note that

$$|X_m'X_m| = |X'X|\,|V_m|(\sigma^2)^{-(k-k_m)} \tag{2.13}$$

where V_m is the variance-covariance matrix of the least squares estimates of the β's omitted by submodel m. Thus, if two competing submodels with the same number of terms yield the same residual sum of squares, the one for which the information about the omitted β's is least (in the sense that the generalized variance $|V_m|$ is greatest) will be least favored.

Remark 2. If all columns of X are multiplied by the same constant c, the effect on (2.10) is to replace γ by $c\gamma$, i.e., the same family of posterior distributions will result. Similarly, multiplying any predictor that is forced into all submodels by a constant c will have no effect. Apart from this, if the relative scales of the columns of X are changed, a different family will result. We shall discuss scaling in more detail in Section 4.1.

From the posterior distribution given in (2.10), one can compute and plot various quantities of interest as functions of γ, where small values of γ reflect a strong prior inclination to include all of the predictors in the model and large values of γ reflect a strong inclination to omit all of the predictors.

When there is interest in evaluating a particular coefficient, β_j, then the following posterior probability is useful:

$$P(\beta_j = 0 \mid y) = \sum P(A_m \mid y) \tag{2.14}$$

where the summation is over all submodels that do not include the coefficient β_j. The posterior probability given in (2.10) can also be used to calculate the posterior expected number of terms in the model:

$$E(k_m \mid y) = \sum_m k_m\, P(A_m \mid y) \tag{2.15}$$

and the posterior entropy:

$$H = -\sum_m P(A_m \mid y)\, ln(P(A_m \mid y)) . \tag{2.16}$$

The entropy is a measure of the degree of dispersion of the posterior probability among the submodels. We find the antilog of the entropy easier to interpret, noting that if the posterior probability were distributed equally among s submodels, the antilog of the entropy would be s.

Plots of (2.14-2.16) as functions of γ are useful in assessing the importance of individual predictors, the number of model terms required, and the extent of uncertainty about the choice of a "best" submodel.

Although we are primarily interested here in the posterior distribution of the submodels, we note that the posterior density of $\boldsymbol{\beta}$ is given by

$$p(\boldsymbol{\beta}\,|\,\mathbf{y}) = \sum_m P(A_m\,|\,\mathbf{y})\,p(\boldsymbol{\beta}\,|\,A_m,\mathbf{y}), \tag{2.17}$$

where $p(\boldsymbol{\beta}\,|\,A_m,\mathbf{y})$ is a multivariate t density centered at $\hat{\boldsymbol{\beta}}_m$ (Box and Tiao (1973), p. 117). In particular, the posterior distribution of β_j is a mixture of scaled and shifted t distributions. The mth such distribution has $n-k_m$ degrees of freedom and is centered at $\hat{\beta}_{m,j}$ with scale factor $q_{m,j}$, where $\hat{\beta}_{m,j}$ is the least squares estimate of β_j in A_m, and $q_{m,j}$ is its standard error.

The posterior distribution of the dependent variable y_p at a specified value $\mathbf{x}_p$ of the predictors is also a mixture of shifted and scaled t-distributions. The mth of these has $n-k_m$ degrees of freedom and is centered at $\mathbf{x}'_{m,p}\hat{\boldsymbol{\beta}}_m$ with scale factor $s_m\sqrt{1+\mathbf{x}'_{m,p}(X_m'X_m)^{-1}\mathbf{x}_{m,p}}$, where $\mathbf{x}_{m,p}$ is the subvector of predictors in $\mathbf{x}_p$ that are present in model A_m and s_m^2 is $S_m^2/(n-k_m)$. These results can be obtained as special cases of results given by Geisser (1965). For fixed γ, the posterior distribution of y_p is a Bayesian predictive distribution (Geisser (1971)). In the next section, we shall suggest a way of using this to assess choices of γ.

2.4 EVALUATION OF γ

2.4.1 Bayesian Cross-Validation

Suppose that we use the Bayesian approach presented here to generate a predictive density $p_{(i)}$ of y_i, the dependent variable in the ith case, given all the data except Y_i, the observed value of the dependent variable in the ith case. All submodels having positive posterior probability will contribute to $p_{(i)}$, i.e., we do not require selection of a "best" submodel first. The goodness of the predictive distribution can be assessed by comparing it to Y_i using some loss function $L(y_i,Y_i)$, which will itself have a distribution generated by the predictive density $p_{(i)}$. Often a single property of the loss distribution, like the mean, can serve as a useful measure of deficiency in the predictive distribution. For squared error loss, the mean of the loss distribution is

$$MSE_i = E(y_i-Y_i)^2 \tag{2.18}$$

where Y_i is known and the expectation E is taken over $p_{(i)}$.

In the spirit of cross-validation, this can be computed for each Y_i in the data set, where the predictive density $p_{(i)}$ in each case is based on all the data except Y_i. We shall take as a measure of the deficiency of prediction the square root of the average of MSE_i over all cases. We shall refer to this as the *predictive error*:

$$PE = \sqrt{\frac{1}{n}\sum_{1}^{n} MSE_i} \ . \tag{2.19}$$

For our variable selection method, PE depends on γ, and can be used as a guide in choosing it. Formulas for the computation of PE are given in Section A.2. It is shown there that PE is finite only if $n - k_m > 3$, so this approach to assessing γ will not be useful unless there are at least four degrees of freedom for error.

Remark 3. In the usual approach to cross-validation, one first defines a *prescription* (Stone (1974)) or *predictive function* (Geisser (1975)) that produces a unique prediction $\hat{y}_{(i)}$ for the ith case given all the data except Y_i. A loss function $L(\hat{y}_{(i)}, Y_i)$ is then defined as a measure of the deficiency of the prescription in the ith case; a common example is the squared error $(\hat{y}_{(i)} - Y_i)^2$. In a Bayesian analysis, $\hat{y}_{(i)}$ would be some function of the predictive distribution $P_{(i)}$; a reasonable choice might be the mean $E_{(i)}$. Under squared error loss, the deficiency of the prescription in the ith case would then be $(E_{(i)} - Y_i)^2$. This would ignore the variance $V_{(i)}$ of the distribution $P_{(i)}$, i.e., all predictive distributions having the same mean would be regarded as equally good. If we try to fix this by weighting the loss by $V_{(i)}^{-1}$ we will overpenalize predictive distributions that are near Y_i and are fairly sharp. We have avoided such difficulties by defining the loss, or deficiency of prediction at a given x_i, directly as a function of Y_i and $P_{(i)}$. That is, we do not first reduce $P_{(i)}$ to a scalar $\hat{y}_{(i)}$. In this respect, our approach is similar to that of Geisser and Eddy (1979), who, in effect, defined the loss to be $-\log p_{(i)}(Y_i)$, where $p_{(i)}$ is the density of $P_{(i)}$.

Remark 4. A plot of PE as a function of γ is useful, since it provides a visual assessment of the effect of γ on the predictive ability of the posterior distribution. Although one could minimize PE as a formal way of choosing γ, we prefer to use the PE plot as an informal guide. Consideration of PE allows us to avoid choosing values of γ that may lead to unacceptably large predictive errors. One should keep in mind, however, that PE is a measure of predictive ability averaged over the cases in the data set at hand. Thus, it is useful when the cases at hand are "representative" of the cases for which one intends to make predictions, but not so useful otherwise.

2.4.2 Goodness of Fit Plot

Another useful way of evaluating γ is to plot the posterior probability of goodness of fit, G, as a function of γ, where G is the sum of the posterior probabilities of all submodels that pass a standard F-test for goodness of fit relative to the full k variable model, at a specified level of significance. For small values of γ, this measure is near 1 since most of the posterior probability is concentrated on the full model. For large values of γ, the measure decreases as more posterior probability is placed on submodels that show significant lack of fit.

This approach to the choice of γ allows us to avoid "Lindley's Paradox," if we so desire. That is, we shall not be in a situation in which high posterior probability is assigned to a model that shows highly significant lack of fit. We should note, however, that it is not always desirable to include terms in a model simply because their exclusion would result in significant lack of fit. Consider, for example, the situation in which $\hat{\beta}_j$ is small enough to have a negligible effect on predictions, yet is significantly different from 0 because its standard error is even smaller. We discuss Lindley's paradox further in Section 4.3. For a more detailed discussion, see Lindley (1957), Shafer (1982), and Smith and Spiegelhalter (1980).

3. EXAMPLES

3.1 BACKGROUND

All of our examples here are based on data from an energy conservation study (Hirst et al., 1985). These data consist of observations on the electricity savings, which is the dependent variable here, for a sample of 401 households that participated in a residential weatherization program. Table 1 contains a list of the ten predictor variables and dependent variable used in this analysis. The first example consists of an analysis of the actual data. In the second and third examples we simulated some of the data to make the analysis more interesting.

Table 1. Variables From Residential Weatherization Program

Dependent Variable	
Y :	Electricity Savings (KWh/year)
Predictor Variables	
X_1:	Presence of air conditioning equipment (0 or 1)
X_2:	Long-run heating degree days
X_3:	Change in number of household members
X_4:	Wood use in 1982-1983
X_5:	Change in electricity price
X_6:	Switched primary heating fuel from electricity (0 or 1)
X_7:	Household income
X_8:	Preprogram electricity use
X_9:	Heated floor area
X_{10}:	Audit prediction of saving

3.2 EXAMPLE 1: THE REAL DATA

The first step in the analysis of this data set was to do the usual regression analyses, using a first order model with 11 terms, including a constant term. (We shall depart slightly from the notation of (1.1) here, in that the dummy predictor associated with the constant term will be denoted by x_0 and its coefficient by β_0.) The results of this analysis included residual plots, calculations of R^2 and Mallows' (1973) C_p for all possible subset regressions, and calculations of regression diagnostics. The residual plots did not show any obvious problems. The magnitude of the variance inflation factors (all < 1.5) and the maximum condition index (2.04) did not imply any significant problems of collinearity among the predictor variables. The full model did not have a high R^2 (0.4652). All of the regression coefficients were significant at the 0.05 level, except for β_1, for which P was 0.497. The submodel that omitted x_1 was the only one for which C_p was less than $1.1k$. (Note: Throughout this section, we shall somewhat arbitrarily regard all submodels that satisfy this criterion as being "acceptable" with respect to C_p.)

We then analyzed these data using Bayesian variable selection as described in Section 2, where all predictors other than the constant term were subject to possible omission from the regression equation. Except for the constant, the original predictors (which we shall denote by $1, X_1, \cdots, X_{10}$) were scaled by subtracting their respective means and dividing by their respective standard deviations. We shall denote these centered and scaled predictors by $x_1, x_2, \ldots, x_{10}$.

The results of the Bayesian analysis included the following plots, all as functions of $ln(\gamma)$:

1. posterior probability that each regression coefficient equals 0 (Fig. 1a);
2. posterior expected number of terms in the model (Fig. 1b);
3. antilog of the posterior entropy (Fig. 1c);
4. predictive error (Fig. 1d);
5. posterior probability of goodness of fit (Fig. 1e).

Figure 1a shows the relative strengths of the individual predictors, the weaker ones rising earlier and increasing to 1 sooner than the stronger ones, as γ increases. We see that the preprogram electricity use (x_8) is a far stronger predictor than the others. Figures 1d and 1e suggest that $ln(\gamma)$ should be less than 6 if we want to minimize our measure of predictive error and maintain a high posterior probability on the submodels that show no lack of fit. For values of $ln(\gamma)$ in this range, the posterior probability is concentrated on the full model, so we would not omit any predictors. Note in Figure 1c that the antilog of the entropy starts to rise sharply at $ln(\gamma) = 6$, showing increasing confusion about the best submodel. For values of $ln(\gamma)$ between 25 and 50, Figures 1a, 1b, and 1c show that the posterior distribution is concentrated on a single submodel, having only two terms, the intercept and $\beta_8 x_8$. Figure 1d shows that this greatly simplified model is acceptable if one is willing to tolerate a predictive error of about 5000 instead of the minimum of 4550. Figure 1e reminds us, however, that this model shows a significant lack of fit.

3.3 EXAMPLE 2: "IMPORTANT" AND "UNIMPORTANT" PREDICTORS

Here we used the same data set, but replaced the values of the dependent variable with values that were generated from a known set of coefficients. The predictors were the same as in Example 1.

The new "observed" values of the dependent variable were generated from:

$$y = f(x) + \epsilon \tag{3.1}$$

where

$$\begin{aligned} f(x) = {} & 3312 + 4755x_1 + 5196x_2 + 455x_3 \\ & + 532x_4 + 538x_5 + 104x_6 + 91x_7 \end{aligned} \tag{3.2}$$

and ϵ represents a normal random variable with mean equal to 0 and a standard deviation of 1000. (We chose the coefficients in (3.2) by selecting β_1 and β_2 randomly from a population of "large" values, β_3, β_4, and β_5 from a population of "moderate" values, and β_6 and β_7 from a population of "small" values; β_8, β_9, and β_{10} were set to zero.)

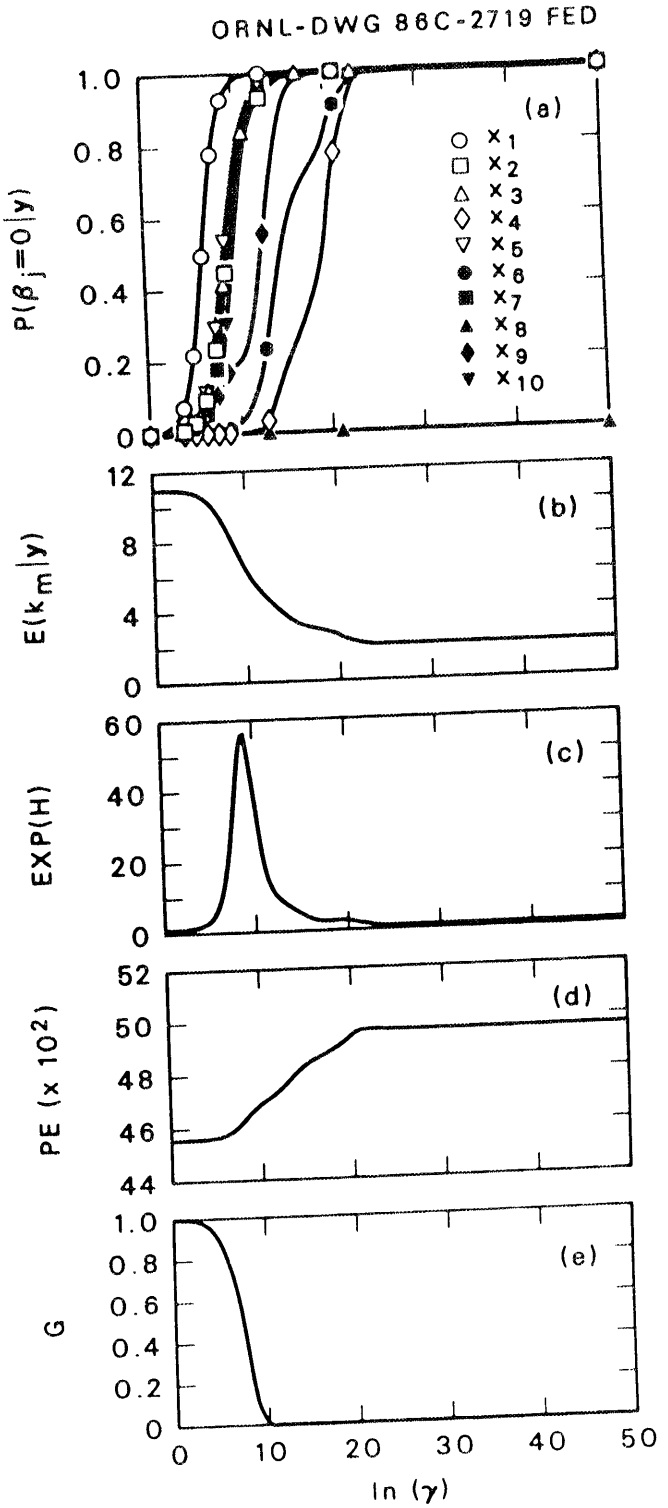

FIGURE 1. Plots of (a) posterior probabilities that each $\beta_j=0$, $p(\beta_j=0|y)$, (b) posterior expected number of terms in model, $E(k_m|y)$, (c) antilog of the posterior entropy, $exp(H)$, (d) predictive error, PE, and (e) posterior probability of goodness of fit, G, as functions of $ln(\gamma)$ for weatherization data with ten predictor variables and observed dependent variable (Table 1).

The least-squares estimate of $f(x)$ was found to be

$$\hat{y} = 3289 + 4712x_1 + 5148x_2 + 440x_3 + 444x_4 + 546x_5 + 51x_6 + 58x_7 - 20x_8 + 85x_9 + 35x_{10}. \tag{3.3}$$

The value of R^2 for the full model is 0.9802. The P-values for Student's t tests of significance of the "small" and "null" β's, β_6–β_{10}, are 0.31, 0.27, 0.73, 0.13, and 0.48, respectively. There are 22 submodels for which C_p is less than $1.1k$. All of these include the terms in x_1–x_5; 11 of them include at least two of the three null terms. The lowest value of C_p is 5.598, for the model that omits the terms in x_6, x_7, x_8, and x_{10}.

The results of the Bayesian procedure are illustrated in Figure 2. Figure 2a sorts out the predictors nicely with respect to strength, although it cannot distinguish between the small ones (x_6,x_7) and the null ones (x_8,x_9,x_{10}). Figures 2d and 2e show that $ln(\gamma)$ can be as great as about 32 without increasing the predictive error or detracting from the goodness of fit. (The predictive error actually decreases slightly from a value of 1377 at $ln(\gamma) = -10$ to a value of 1372 at $ln(\gamma) = 25$.) Figures 2b and 2c show that for $ln(\gamma)$ about 32, the posterior probability is concentrated on a single model having 6 terms; Figure 2a shows that this model contains the predictors x_1–x_5. This is clearly the model of choice, unless one were willing to tolerate an increase in predictive error to about 1700 (Figure 2d) in order to obtain an even simpler model. In this case, one would choose $ln(\gamma)$ greater than about 37, where the posterior probability is concentrated on the model containing predictors x_1 and x_2.

3.4 EXAMPLE 3: COLLINEARITY

In this example, strong collinearity was introduced among some of the predictor variables. This was done by replacing the predictor X_9 by

$$X_{9,NEW} = \hat{X}_9 + 0.01(X_9 - \hat{X}_9) \tag{3.4}$$

where $\hat{X}_9$ is the least squares predicted value of X_9 when it is regressed on X_1, X_3, X_6, X_7, and X_8 with a constant term included in the regression equation. A second collinear relation was introduced by replacing the predictor X_2 by

$$X_{2,NEW} = \hat{X}_2 + 0.01(X_2 - \hat{X}_2) \tag{3.5}$$

where $\hat{X}_2$ is the least squares predicted value of X_2 when it is regressed on X_1, X_3, X_4, X_6, and X_7 with a constant included in the regression equation. Before the "observed" values of the dependent variable (y) were generated, the ten predictor variables were centered and scaled using the same constants used in the previous examples. The generated values of the dependent variable were obtained using the same β's and ϵ's used in Example 2. Thus the only difference between Examples 2 and 3 is in the data for predictors x_2 and x_9. The "true" response function, the vector of errors, and the prior distribution of the β's are unchanged.

The least squares estimates of the regression coefficients, the P-values for the associated t-statistics, and the variance inflation factors are given in the following table.

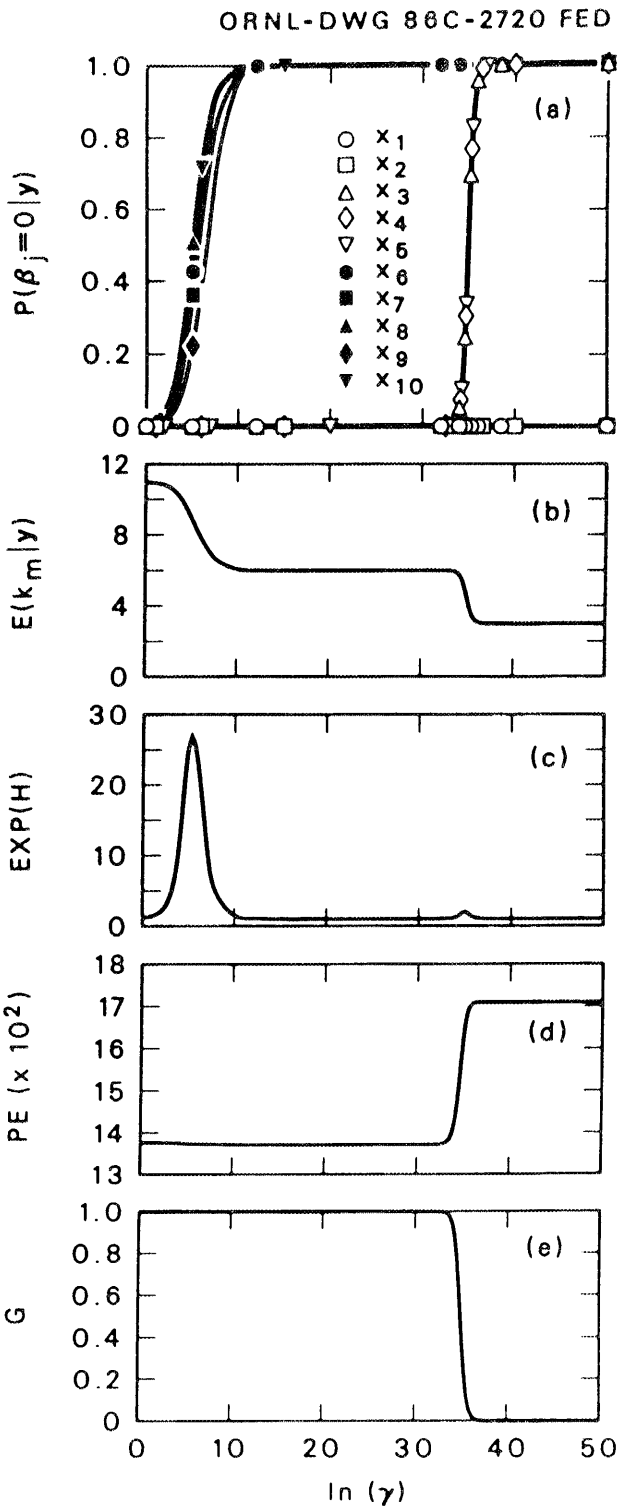

FIGURE 2. Plots of (a) posterior probabilities that each $\beta_j=0$, $p(\beta_j=0|y)$, (b) posterior expected number of terms in model, $E(k_m|y)$, (c) antilog of the posterior entropy, $exp(H)$, (d) predictive error, PE, and (e) posterior probability of goodness of fit, G, as functions of $ln(\gamma)$ for weatherization data with ten predictor variables and simulated values of dependent variable (Section 3.3).

j	β_j	$\hat{\beta}_j$	P	VIF
1	4755	3863	<0.01	123.0
2	5196	338	0.95	263.5
3	455	220	0.47	40.6
4	532	824	0.04	70.4
5	538	546	<0.01	1.2
6	104	-455	0.37	112.1
7	91	-1968	0.16	832.2
8	0	-2964	0.13	1679.8
9	0	8507	0.13	3143.1
10	0	35	0.48	1.1

The effects of the collinearity are evident. As one might expect, this leads to considerable confusion regarding the choice of a "best" model. The value of R^2 for the full model is 0.952; there are 89 submodels having values of R^2 greater than 0.950. There are 13 submodels having values of C_p less than $1.1k$; these all include the terms in x_1, x_4, x_5, and x_7. Six of the thirteen exclude x_2, and eight of them include at least two of the three null terms. The lowest C_p is 5.87, for the model that includes x_1, x_3, x_4, x_5, x_6, and x_7.

The results of the Bayesian analysis are illustrated in Figure 3. It appears that one can choose $ln(\gamma)$ as great as about 10 without degrading goodness of fit (Figure 3e) or predictive capability (Figure 3d). At this point, the posterior distribution is spread over several submodels (the antilog of the entropy being about 4 in Figure 3c). Figure 3a suggests that at $ln(\gamma) = 10$ the terms in x_7, x_8, x_9, and x_{10} can be omitted, and the rest should probably be kept, though there is some doubt about x_4 and x_6. An examination of the posterior probabilities of the submodels when $ln(\gamma) = 10$ reveals that the two most favored submodels are

$$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_6x_6,$$

which has probability 0.48, and

$$\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5,$$

which has probability 0.38. The next most favored has probability 0.02.

If one does not mind lack of fit, but is concerned mainly with predictive error, larger values of $ln(\gamma)$ (up to about 40) are acceptable. At this point, all of the posterior probability is concentrated on the model that includes x_1, x_4, x_5, and x_6.

The non-monotone behavior of some of the curves in Figure 3a is obviously caused by the collinearity, but we have not investigated the relationship.

4. DISCUSSION

4.1 CENTERING AND SCALING

In general, the predictors x_j used in the analysis are the result of "centering and scaling" the original predictors X_j, i.e.,

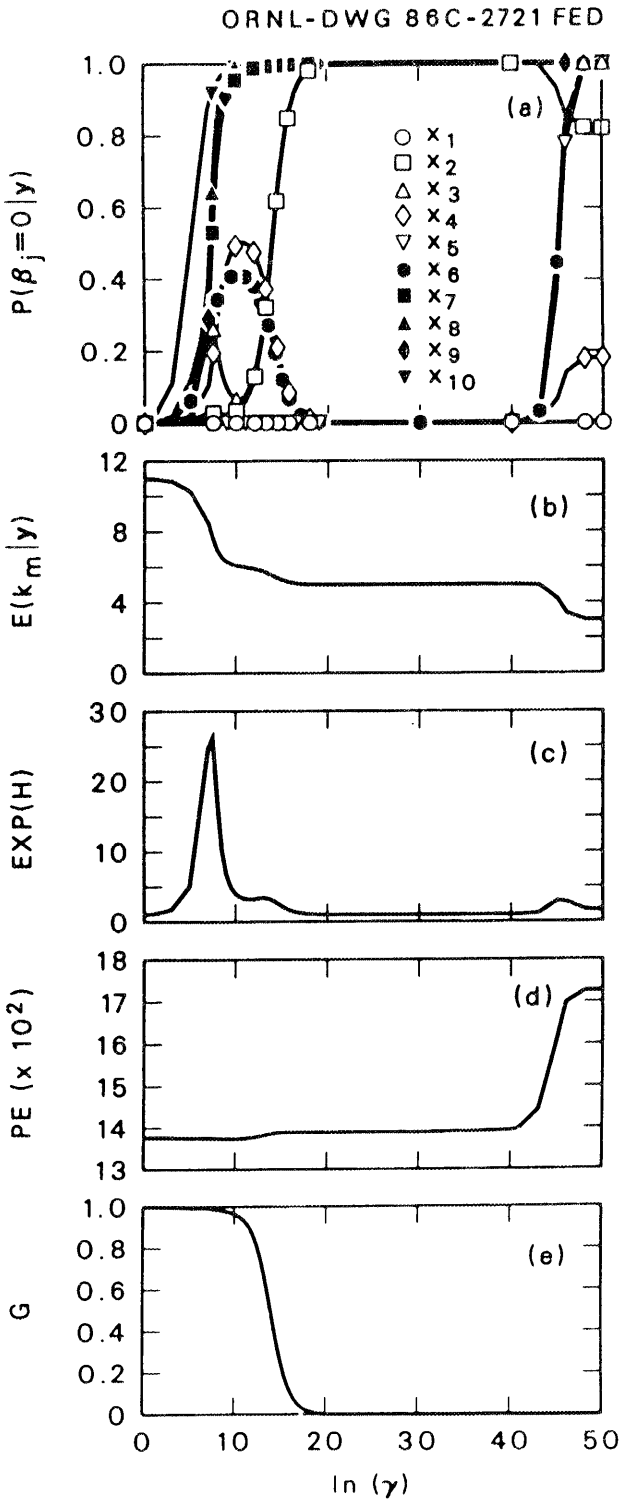

FIGURE 3. Plots of (a) posterior probabilities that each $\beta_j=0$, $p(\beta_j=0|y)$, (b) posterior expected number of terms in model, $E(k_m|y)$, (c) antilog of the posterior entropy, $exp(H)$, (d) predictive error, PE, and (e) posterior probability of goodness of fit, G, as functions of $ln(\gamma)$ for weatherization data with ten predictor variables and simulated values of dependent variable and collinear predictor variables (Section 3.4).

$$x_j = (X_j - c_j)/\, d_j \tag{4.1}$$

where c_j is the "centering constant" and d_j is the "scaling constant" for the jth predictor. Here we discuss the effect of these constants on the Bayesian analysis described in Section 2, and offer some suggestions for choosing them.

The choice of centering constants will rarely cause any problems. Usually, there is a constant term in the model, and it is exempt from deletion; then any choice of centering constants will do, since the posterior probabilities in (2.9) will be unaffected. Occasionally, one may want to fix $E(y)=\mu_0$ at some point $X_{10}, X_{20}, \cdots, X_{k0}$. In this case, take $c_j = X_{j0}$, replace y by $y-\mu_0$ and omit the constant term from the model equation.

When the values of γ_j in (2.9) are "known," through prior specification of f_j and h_{0j} in (2.4), scaling doesn't matter, since any rescaling would be accompanied by a suitable modification of the f's and the γ's. This is not the case, however, for the approach we developed following (2.9) to avoid having to specify the γ's *a priori*. This approach, which uses (2.10), requires identical prior distributions for all β's that are subject to omission from the regression equation. If we adhere to this, then as noted in Remark 2 in Section 2.3, a change in the relative scales of the columns of X essentially changes the prior distribution, resulting in a different family of posterior distributions. Our intent in choosing identical priors was to reflect prior impartiality about the relative "importance" of the predictors. We propose that the "importance" of the jth predictor be defined in some reasonable way as a multiple of $|\theta_j|$, where θ_j is the coefficient of the original predictor X_j, and that the scaling constant d_j be set equal to the multipler for each j. The choice of identical priors for the β's then implies that the prior distribution of importance is the same for all predictors. Note that this requires no prior information about the values of the θ's.

We suggest the following two types of scaling based on this approach.

Type 1. For each raw predictor X_j, choose a "range of prediction," an interval of width r_j that covers the values of X_j for which one expects to use the regression equation. Define the "importance" of X_j to be $|\theta_j r_j|$, the absolute change in $E(y)$ effected by moving X_j across its range of prediction. Now set the scaling constant d_j equal to r_j; this will make the importance of the jth predictor equal to $|\beta_j|$.

Type 2. Another definition of importance depends on the *distribution* of future predictor values. Suppose that the future distribution of values of X_j at which one expects to use the regression equation has a standard deviation σ_j. Then the standard deviation of the contribution $\theta_j X_j$ to the prediction of y is $|\theta_j \sigma_j|$; define the importance of the jth predictor to be equal to this. Now choose the scaling constant d_j equal to σ_j; this will make the importance of the jth predictor equal to $|\beta_j|$. If the values of X_j in the data at hand are "representative" of the population for which one intends to make future predictions, then one can use the standard deviation of X_j computed from the observed data to estimate σ_j. This justifies the "standard" scaling used in our examples.

Note, however, that the rationale for both types of scaling is much weaker if the values of the predictors at which one would like to make predictions are correlated. In this case, it is much more difficult to define the "importance" of a predictor, and hence to justify a particular choice of scaling constants. The same remark applies to models in which some of the predictors are known functions of others, e.g., a quadratic response surface model. More attention to this point is needed, we think, before the methods proposed in this paper can be applied to such models.

4.2 INTERPRETATION OF BAYESIAN PROBABILITIES

When the regression coefficients are regarded not as properties of the real world, but as part of a mathematical device (the model) that has been made up for the purpose of prediction, we find the following approach helpful in defining the meaning of probability statements about them.

We envision a hypothetical population J of "experts," each of whom can produce, *a priori*, a predicted value of the dependent variable y in any case given the values of the predictors in that case. Each of these experts bases all of his predictions on a specific value of β and σ. In a single case, a prediction is made using (1.1), where ϵ is drawn randomly from a normal distribution with mean 0 and standard deviation σ. In multiple cases, these drawings are made independently.

Prior probabilities are defined in terms of the distribution of choices of the model parameters within the population J. For example, the prior probability that $\beta_j = 0$ is the proportion of experts in J who choose to omit x_j from the model. We are defining probability here as frequency within a hypothetical population of experts, in contrast to the "classical" statistical models in which probability is defined as frequency within a hypothetical population of similar experiments.

Having taken n observations on the predictors and the dependent variable, suppose we remove from J all of the experts who did not predict all of the y's correctly (to within an arbitrarily small tolerance). The posterior distribution of the β's and σ, as obtained from Bayes' Theorem, can be regarded as the distribution of these parameters within the population J' of surviving experts.

This approach does not really require an expression of prior belief by the user, although it is unlikely that one would use it if one disapproved of the prior makeup of J. To interpret probabilities here as a measure of personal belief, one would need to assert that, in any "thought experiment," the joint distribution of y implied by the prior distribution on the model parameters represents one's prior uncertainty about y. Even if it were possible to do this in practice, it would still be difficult to explain what is meant by probability statements that involve the model parameters, which are artificial constructs.

We have deliberately tried to choose a Bayesian approach in which the prior beliefs of the user are as unobtrusive as possible. The prior marginal distributions of the β's are all the same, which, under the scaling conventions suggested above, should make the procedure fair to all predictors. Moreover, given a particular submodel, the β's that appear in that submodel have the most commonly accepted "noninformative" joint prior. Finally, the user need not specify in advance his strength of belief that any particular predictor is null, since γ is an adjustable constant that can be assessed using the data (Section 2.4).

Our approach can be modified, if desired, to give the user more control over the prior distribution without having to specify all of its parameters explicitly. As before, we assume the predictors have been suitably scaled. Set all the f_j's in (2.3) to be equal to an unspecified constant; then choose the "odds for deletion" $\phi_j = h_{0j}/(1 - h_{0j})$ for each predictor. By (2.4), $\gamma_j = \gamma\phi_j$, where γ is an unspecified constant. Now use (2.9), and evaluate the parameter γ as described in Section 2.4. Note that it is only the relative odds for deletion that matter, since if all the ϕ's are multiplied by a constant, the same family of posterior distributions arises.

4.3 A COUPLE OF PARADOXES

Since our Bayesian model places probability at a specific point (0) in the distribution of any β that is a candidate for omission from the regression equation, we should expect "Lindley's paradox" to occur. In fact, the Bayesian model used by Lindley (1957) to demonstrate the paradox is the same as ours, restricted to one dimension and with σ^2 known. In our setting, Lindley's paradox is that, given any value of γ, there are data sets such that a submodel can show significant lack of fit based on the usual F-test and can also have posterior probability close to 1. This occurs when, for example, $\hat{\beta}_j$ is close to 0 yet its standard error is small enough to make it significantly different from 0. This phenomenon, which was first noted by Jeffreys (1967), has caused considerable discussion in the literature. We have little to add, except to explain why we are not worried about it here.

In the context of the population of predicting experts J described in Section 4.2, the significance of $\hat{\beta}_j$ at the 0.01 level, for example, means that

(A) among those experts who omitted x_j from the model, less than 1% predicted that the lack of fit statistic would be greater than the value that was actually observed.

At the same time (if Lindley's paradox occurs) the posterior probability that $\beta_j = 0$ may be close to 1, indicating that

(B) the proportion of experts who correctly predicted y was far greater among those who omitted x_j from the model than among those who included this term but were not in agreement about the value of β_j.

In this context there is no paradox, since there is no particular reason why (A) should necessarily be consistent with (B). In the first place, the two events are different, as Lindley (1957) pointed out. Secondly, (A) refers only to those experts who omit the term in x_j, while (B) considers their performance relative to those who include it.

We think that statements like (B) are usually more relevant to the variable selection problem than are statements like (A). In any case, we note that our "goodness of fit" assessment of the choice of γ (Section 2.4) allows us to avoid, if we want, values of γ that would result in a high posterior probability for submodels that showed a significant lack of fit by the classical test. For further discussion of Lindley's paradox, see Lindley (1957), Bartlett (1957), Shafer (1982), and Smith and Spiegelhalter (1980).

While on the subject of paradoxes, note that if we fix h_0 (the prior probability that $\beta_j = 0$), then as f (the half-width of the interval on which each β_j has positive prior probability) approaches ∞, the posterior probability concentrates totally on the submodel having the fewest possible predictors no matter what the data. (This can be seen easily from equation (2.10), since γ approaches ∞.) This was noted by Bartlett (1957) in the one-dimensional example used by Lindley (1957), and has been the subject of more extensive investigation by Atkinson (1978) and Pericchi (1984).

The difficulty can be discussed most easily in one dimension. Here the prior probability that β is in any interval that does not include 0 approaches 0 as f approaches ∞, while the prior probability that β is in any interval that includes 0 approaches a constant. We have avoided this difficulty by introducing the parameter γ; in the limiting process,

γ is held fixed so h_0 approaches 0 as f approaches ∞. This was suggested by Lindley in correspondence with Bartlett, but Bartlett (1957) regarded it as "rather an artificial evasion of the difficulty." To us it seems a sensible evasion of the difficulty; it is the difficulty itself that is artificial. This can be seen by observing that, as long as f is large enough for (2.7) to hold, the effect of the prior distribution on the posterior distribution is a function of γ only. Thus, a whole family of posterior distributions, indexed by γ, can be generated. Letting f approach ∞ while h_0 is fixed drives the posterior distribution toward the one member of this family that corresponds to $\gamma=\infty$. This seems to us unnecessarily restrictive.

4.4 OTHER METHODS

Box and Meyer (1986) developed a Bayesian variable selection approach for the problem of identifying the important effects in unreplicated fractional factorial designs. They supposed that each effect β_j is "active" (non-zero) with probability α, and that all effects are independently and identically distributed. Our approach, in the somewhat more general linear regression setting, stems from the same idea, but differs in some major respects. We specify a locally uniform prior for each active β_j, while Box and Meyer use a normal prior, the variance of which is a "known" multiple q of σ^2, where $q=k^2-1$, (k being a known parameter here, not the number of model terms). Their posterior distributions depend on two parameters, α and k, which are supplied by the user. Our posterior distributions depend on one parameter, γ, which is treated as an adjustable constant and assessed as suggested in Section 2.4.

In the Box-Meyer model, the prior variance of each active β_j is specified to be a multiple of σ^2. This practice, which greatly simplifies the derivation of the posterior distributions, seems to be adopted without question by most authors. One of its consequences is that one always has n degrees of freedom for the estimation of σ^2, no matter how many terms there are in the regression equation. In the Box-Meyer setting, for example, if all effects are known to be active, the inference about σ^2 is the same as if $\hat{\beta}_1 k^{-1}, \hat{\beta}_2 k^{-1}, \ldots$, were treated as a sample from a normal distribution with known mean 0. If there is no link between the variances of the β's and σ^2, then the data provide no information about σ^2 when the rank of X is n. We have chosen not to provide such a link, so there are some restrictions on the maximum rank of X if one is to apply the methods of Section 2.4 to assess values of γ.

Several closely related Bayesian variable selection methods have been considered by other authors (Atkinson (1978), Halpern (1973), Pericchi (1984)), from the viewpoint of model selection or model discrimination. This is a somewhat more general framework than the one set up by Box and Meyer and by us. As a point of reference, consider the following "standard" setup (Pericchi (1984)), in which the components of the prior distribution are as follows:

$p(\boldsymbol{\beta}_m \mid A_m, \sigma)$ is a multinormal density with mean $\boldsymbol{\beta}^*_m$ and covariance matrix $\sigma^2 \boldsymbol{V}_m$.

$p(\sigma^{-2} \mid A_m)$ is a two-parameter gamma density whose parameters may depend on m.

$P(A_m)$ is specified by the user.

If the user can supply the needed parameters $P(A_m)$, $\boldsymbol{\beta}^*_m$, and $\boldsymbol{V}_m$ for all m, as well as the parameters of the gamma distributions for σ^{-2}, the analysis is straightforward. (See Pericchi (1984) for the formulas needed to compute the posterior probabilities of the

competing models.) To relieve the user of the burden of specifying all these parameters (and the fear of having to defend his choices), various simplifications can be adopted to express ignorance or impartiality. For example, one might choose $\beta^*_m=0$ and $V_m=qI$, where q is a parameter that can be specified or treated as an adjustable constant. One might also choose the standard noninformative prior distribution for σ, and thus avoid having to choose the parameters of the gamma priors for σ^{-2}. None of these simplifications cause any special difficulties, and can be used whenever they seem sensible. Unfortunately, seemingly sensible attempts to express impartiality toward the competing models and prior ignorance about the magnitudes of the β's within those models have led to difficulties of the type exemplified by Bartlett's paradox. (See Atkinson (1978) for a more extensive discussion of these difficulties.)

Pericchi (1984) argued that the prior submodel probabilities should be adjusted by the factor $exp(I_m)$, where I_m is the expected gain in information (defined as the expected change in entropy) in the distribution of the submodel parameters. The effect of this is to remove from the posterior distribution of the submodels those terms that are responsible for paradoxes of the kind described Section 4.3. One unsatisfying aspect of this is that the prior distribution of the submodels becomes dependent on the design. Moreover, the effect of the design on the posterior probabilities seems to us to be in the wrong direction. Consider two competing models, one containing only the predictor x_1 and the other containing only the predictor x_2. If the design is good for estimating β_1 but bad for estimating β_2, the first model will be favored, *a priori*, by Pericchi's rule. That is, the expectation of weak information about β_2 will increase the probability that β_2 is declared to be 0. To our minds, this presents more of a difficulty than does Lindley's paradox, which Pericchi's rule effectively banishes.

There are, of course, numerous non-Bayesian techniques for variable selection. See, e.g., the review by Hocking (1976). Classical approaches have produced many useful statistics, e.g., Mallows' (1973) C_p, Allen's (1971b) PRESS, P-values for specified lack-of-fit tests, and so forth, which the experienced statistician considers in arriving at a reasoned opinion regarding the consequences of omitting predictors from the model. However, most classical methods are based on statements of the form: "If the data are generated by submodel A_m, then the probability of the event E (or the expectation of the statistic Z) is Q." There are a great many potentially informative statements of this type, corresponding to different choices of A_m and E or Z. The process of synthesizing this information and resolving the inevitable conflicts is not an easy one, especially when data-guided strategies change the Q's in unknown ways. For a good exposition of the difficulties, see the paper by Miller (1984).

In contrast, the process of synthesizing the available information using a Bayesian approach is relatively straightforward, subject to computational limitations. All of the agony centers on the choice of prior distribution(s). We have simplified further by reducing this agony to the choice of a single parameter γ, in addition to the choice of scaling. We have suggested, in Sections 4.1 and 2.4, some ways to make these choices. Although it might be argued that we have oversimplified by restricting attention to too narrow a class of priors, we think that this class is flexible enough to be of practical use.

Another feature we find appealing about a Bayesian approach to the problem of selecting predictors is that the results are produced in the form of probabilities for the various submodels. This is unlikely to impress those who think that Bayesian probabilities are of questionable value, although we hope that our attempt in Section 4.2 to endow these probabilities with a frequentist meaning will help.

We think that the major practical limitations of the methods we have presented here are the exclusion of functional dependencies among predictors, the computational effort required if there are many predictors, and the restriction of the error model to the normal distribution. We hope that these limitations will be removed as a result of further work in this area, and that, in meantime, the methods we have described here will be a useful adjunct to methods of variable selection that are currently in use.

APPENDIX

Here we present some formulas and computational details that we have used in carrying out the methods described in this paper.

A.1 NOTATION

Consider a single regression model, where the data is in the form of an $n \times k$ matrix X and an n-vector y. The following defines the necessary notation:

x_i' is the ith row of X

$$c_i = (X'X)^{-1} x_i$$

$$h_i = x_i'(X'X)^{-1} x_i = c_i' x_i$$

$$b = (X'X)^{-1} X'y$$

$$\hat{y}_i = x_i' b = c_i' X'y$$

$$e_i = y_i - \hat{y}_i$$

$$f_i = e_i / (1 - h_i)$$

$$S^2 = \sum_1^n e_i^2$$

$$s^2 = S^2 / (n - k)$$

A.2 PREDICTIVE ERROR FOR THE ith CASE

For the moment, assume there is only one regression model, with k terms. Let $X(i)$, $y(i)$, $c_i(i)$, $h_i(i)$, $b(i)$, $\hat{y}_i(i)$, $e_i(i)$, $S^2(i)$, $s^2(i)$ be defined as above when the ith case is omitted from the data set. The following formulas are useful in computing the effect of omitting a case:

$$h_i(i) = h_i / (1 - h_i) \tag{A.2.1}$$

$$b(i) = b - f_i c_i \tag{A.2.2}$$

$$\hat{y}_i(i) = \hat{y}_i - h_i f_i \tag{A.2.3}$$

$$S^2(i) = S^2 - e_i f_i \tag{A.2.4}$$

$$s^2(i) = (S^2 - e_i f_i) / (n - k - 1) \tag{A.2.5}$$

$$det\,[X'(i) X(i)]^{-1} = det\,[X'X]^{-1} / (1 - h_i) \tag{A.2.6}$$

The posterior distribution of the dependent variable in the ith case given all the data other than y_i is a scaled and shifted t-distribution, centered at $\hat{y}_i(i)$ with scale factor $s(i)\sqrt{1 + h_i(i)}$ and $n - k_i - 1$ degrees of freedom.

The mean squared predictive error for the ith case is

$$MSE_i = (y_i - \hat{y}_i(i))^2 + s^2(i)(1+h_i(i))(\frac{n-k-1}{n-k-3}), \tag{A.2.7}$$

the last factor on the right being the variance of the t-distribution with $n-k-1$ degrees of freedom.

Substituting (A.2.1, A.2.3, and A.2.5) into (A.2.7) yields

$$MSE_i = f_i^2 + (S^2 - e_i f_i)(1-h_i)^{-1}(n-k-3)^{-1} \ . \tag{A.2.8}$$

This pertains to a single model with k terms. When there are several models, indexed by $m=1,2,\ldots$, it can be shown that MSE_i is a weighted average of terms of the form (A.2.8), i.e.,

$$MSE_i = \sum_m P_{(i)}(A_m)[f_{i,m}^2 + (S_m^2 - e_{i,m} f_{i,m})(1-h_{i,m})^{-1}(n-k_m-3)^{-1}], \tag{A.2.9}$$

where $P_{(i)}(A_m)$ is the posterior probability of model m given all the data except y_i.

Our summary measure (PE in equation (2.19)) of the predictive error is the square root of the average of MSE_i over all n cases.

A.3 COMPUTATION

To calculate the posterior probabilities of the submodels, one needs to compute the residual sum of squares S_m^2 and $|X_m'X_m|$ for every submodel that has been assigned a positive prior probability.

If one wants to compute the predictive error PE, more computational work is required. Most of this involves the computation of the vectors $\boldsymbol{h}_m$ and $\boldsymbol{e}_m$ for each submodel. One's approach will depend on the numerical software available; here we describe the main components of our computer program, which is written in FORTRAN 77 and makes use of the NAG subroutine library.

Given n, k_m, X_m, y, $X_m'X_m$, and $X_m'y$, the main subroutine first does a Cholesky decomposition of $X_m'X_m$, using the NAG routine F01BQF. That is, U and D are found such that $X_m'X_m = U'DU$, where U is a unit upper triangular matrix and D is a diagonal matrix with diagonal $\boldsymbol{d}$.

The determinant of $X_m'X_m$ is just the product of the d_j's.

The inverse of U, which is the unit upper triangular matrix Q, is computed using

$$Q_{i,j} = -U_{i,j} - \sum_{k=i+1}^{j-1} U_{i,k} Q_{k,j} \qquad i<j. \tag{A.3.1}$$

(The summation on the right is omitted if $j=i+1$.)

For each case i, let $z_i = Q'x_i$. Further, let $w = Q'X'y$. The residuals are obtained from $e_i = y_i - \hat{y}_i$, where $\hat{y}_i = \sum_{j=1}^{k} z_{ij} w_j / d_j$. S^2 is then computed directly from the residuals. The elements of the hat vector h are obtained from $h_i = \sum_{j=1}^{k} z_{ij}^2 / d_j$. The elements of the mean squared predictive error vector for the submodel are given by equation (A.2.8). The log model weights w_m, as given by equation (2.12), are also calculated and returned by the subroutine.

To save storage space, the predictive mean squared error vector, whose i^{th} element is given by (A.2.9), is accumulated as a running weighted average and is updated after the calculations for each submodel have been made.

When all possible submodels have been considered, the posterior distribution $P(A_m \mid y)$ is obtained from the log submodel weights, and the posterior entropy is computed. Other properties of the posterior distribution, e.g., $P(\beta_j = 0 \mid y)$, can now be easily obtained.

Remark A1. Once $ln(w_m)$ in (2.12) has been computed for one value of γ, the computation for other values of γ is immediate. This should be exploited in the program.

Remark A2. Our current program does the main computations for each submodel independently. We expect that our current program could benefit from a more efficient approach, under which computations made for one submodel can be used in obtaining the results for another. See, for example, Furnival and Wilson (1974).

ACKNOWLEDGEMENT

We would like to express our appreciation to E. Hirst and D. White of the Energy Division at Oak Ridge National Laboratory for the weatherization program data, and to M. D. Morris of the Engineering Physics and Mathematics Division at Oak Ridge National Laboratory for constructive comments on an early draft of this manuscript.

REFERENCES

[1] Allen, D.M. (1971a), "Mean Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics*, 13, 469-475; discussion, 477-481.

[2] Allen, D.M. (1971b), "The Prediction Sum of Squares as a Criterion for Selecting Variables," *Technical Report No. 23*, Department of Statistics, University of Kentucky.

[3] Atkinson, A.C. (1978), "Posterior Probabilities for Choosing a Regression Model," *Biometrika*, 65, 39-48.

[4] Bartlett, M.S. (1957), "A Comment on D.V. Lindley's Statistical Paradox," *Biometrika*, 44, 533-534.

[5] Box, G.E.P. and Meyer, R.D. (1986), "An Analysis for Unreplicated Fractional Factorials," *Technometrics*, 28, 11-18.

[6] Box, G.E.P. and Tiao, G.C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

[7] Draper, N.R. and Smith, H. (1981), *Applied Regression Analysis, Second Edition*, New York: Wiley.

[8] Furnival, G.M. and Wilson, R.W., Jr. (1974), "Regressions by Leaps and Bounds," *Technometrics*, 16, 499-511.

[9] Geisser, S. (1965), "Bayesian Estimation in Multivariate Analysis," *Ann. Math. Statist.*, 36, 150-159.

[10] Geisser, S. (1971), "The Inferential Use of Predictive Distributions," in *Foundations of Statistical Inference*, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, 456-469.

[11] Geisser, S. (1975), "The Predictive Sample Reuse Method with Applications," *Journal of the American Statistical Association*, 70, 320-328.

[12] Geisser, S. and Eddy, W.F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153-160; correction 75, 765.

[13] Halpern, E.F. (1973), "Polynomial Regression from a Bayesian Approach," *Journal of the American Statistical Association*, 68, 137-143.

[14] Hocking, R.R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-51.

[15] Hirst, E., White, D., Holub, E.,and Goeltz, R. (1985), "Actual Electricity Savings for Homes Retrofit by the BPA Residential Weatherization Program," Oak Ridge National Laboratory Report ORNL/CON-185.

[16] Jeffreys, H. (1967), *Theory of Probability* (3rd. ed.), Oxford: Oxford University Press.

[17] Lindley, D.V. (1957), "A Statistical Paradox," *Biometrika*, 44, 187-192. (Comments in Vol. 45, 533-534.)

[18] Mallows, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661-675.

[19] Miller, A.J. (1984), "Selection of Subsets of Regression Variables," *Journal of the Royal Statistical Society, Series A*, 147, 389-425.

[20] Pericchi, L.R. (1984), "An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models," *Biometrika*, 71, 575-586.

[21] Shafer, G. (1982), "Lindley's Paradox," *Journal of the American Statistical Association*, 77, 325-334. (Comments in Vol. 77, 334-351.)

[22] Smith, A.F.M. and Spiegelhalter, D.J. (1980), "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society, Series B*, 42, 213-220.

[23] Stone, M. (1974), "Cross-validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Series B*, 36, 111-147.

DISTRIBUTION OF ORNL-6328

1. C. K. Bayne
2-6. J. J. Beauchamp
7. K. O. Bowman
8. M. V. Denson
9. D. T. Downing
10. E. L. Frome
11. D. G. Gosslee
12. L. J. Gray
13-14. Rhonda Harbison/ Mathematical Sciences Library
15. M. T. Heath
16. T. L. Hebble
17. H. Hwang
18. J. K. Ingersoll
19. W. F. Lawkins
20. W. E. Lever
21. F. C. Maienschein
22-26. T. J. Mitchell
27. M. D. Morris
28. G. Ostrouchov
29. V. R. R. Uppuluri
30. R. C. Ward
31. C. Weisbin
32. D. White
33. D. G. Wilson
34. T. Wright
35. A. Zucker
36. Central Research Library
37. K-25 Plant Library
38. ORNL Patent Office
39. Y-12 Technical Library Document Reference Section
40. Laboratory Records - RC
41-42. Laboratory Records Depatment
43. P. W. Dickson, Jr. (Consultant)
44. G. H. Golub (Consultant)
45. R. M. Haralick (Consultant)
46. D. Steiner (Consultant)

EXTERNAL DISTRIBUTION

47. Dr. Donald M. Austin, ER-7, Applied Mathematical Sciences, Scientific Computing Staff, Office of Energy Research, Office G-437 Germantown, Washington, DC 20545.

48. Dr. Herman Chernoff, Statistics Center, Massachusetts Institute of Technology, Cambridge, MA 02139.

49. Dr. N. R. Draper, Department of Statistics, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706.

50. Dr. Jerome H. Friedman, Department of Statistics, Stanford University, Stanford, CA 94305.

51. Dr. Irwin Guttman, Department of Statistics, University of Toronto, Toronto, Ontario, CANADA M5S 1A1.

52. Dr. Eric Hirst, 0BC-06S, Puget Sound Power and Light Co., P.O. Box 97034, Bellevue, WA 98009.

53. Dr. Ronald Hocking, Department of Statistics, Texas A&M University, College Station, TX 77843.

54. Dr. C. Nachtsheim, School of Management, University of Minnesota, Minneapolis, MN 55455.

55. Dr. Ronald Peierls, Applied Mathematics Department, Brookhaven National Laboratory, Upton, NY 11973.

56. Dr. Darius Sabavala, Bell Communications Research, 290 West Mt. Pleasant Avenue, Livingston, NJ 07039-2729.

57. Office of Assistant Manager for Energy Research and Development, Department of Energy, Oak Ridge Operations Office, Oak Ridge, TN 37830.

58-87. Technical Information Center.